AD________________

Award Number: W81XWH-10-1-0463

TITLE: Origins of DNA Replication and Amplification in the Breast Cancer Genome

PRINCIPAL INVESTIGATOR: Susan A. Gerbi, Ph.D.
Alexander Brodsky, Ph.D.
Ben Raphael, Ph.D.

CONTRACTING ORGANIZATION: Brown University
Providence, RI 02912

REPORT DATE: September 2011

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE
September 2011

2. REPORT TYPE
Annual

3. DATES COVERED
1 September 2010 – 31 August 2011

4. TITLE AND SUBTITLE

Origins of DNA Replication and Amplification in the Breast Cancer Genome

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-10-1-0463

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Susan A. Gerbi
Alexander Brodsky
Ben Raphael
E-Mail: Susan_Gerbi@Brown.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Brown University
Providence, RI 02912

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The major goal of this IDEA Expansion Breast Cancer Research grant is to test whether a correlation existsbetween sites of DNA amplification and estrogen receptor (ER) binding in the breast cancer genome. Correlations would support our hypothesis that ER adjacent to replication origins may interact with the replication machinery to drive DNA amplification, a hallmark of many cancers. Our Specific Aims are:
(1) Map replication origins in the MCF-7 breast cancer genome by genomic sequencing
(2) Compare the replication origin maps between breast cancer (ER+, ER-) and normal breast cells
(3) Correlate the origin map data with sites of DNA amplification and estrogen receptor binding
(4) Pilot runs to map replication origins in ER+ human breast cancer tissue and sites of DNA amplification

15. SUBJECT TERMS
estrogen receptor, DNA amplification, replication origins

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
					USAMRMC
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(include area code)*
U	U	U	UU	21	

Table of Contents

Introduction

Genetic instability and rearrangements, including gene amplification, is a hallmark of many cancers. It would be desirable to prevent gene amplification, thereby moderating the aggressive growth of breast cancer cells. The problem is that no one knows what triggers gene amplification. Our current research in a model system suggests that the trigger may be a transcription factor such as the receptor for the steroid hormone estrogen. The research proposed here describes some experiments to begin to test this idea. Cancer is believed to occur after a build-up of somatic mutations or other genomic changes. We wish to ask if a genomic change (genetic or epigenetic) might produce novel binding site(s) for the estrogen receptor (ER) near a replication origin to cause re-replication, resulting in amplification. In order to address this question, as a first step we need to map all origins of replication in the human (breast cancer) genome --- which is the **subject** of this DOD-funded grant. The **purpose and goal** of these experiments is to be able to see if correlations exist between origins that re-replicate (leading to DNA amplification) in breast cancer cells and sites of estrogen receptor binding. The **scope** of our research is summarized by the three Specific Aims: (1) Map replication origins in the human genome; (2) Comparison of replication origin maps between breast cancer (ER+, ER-) and normal breast cells; (3) Correlation of origin map data with sites of (a) DNA amplification and (b) estrogen receptor binding. The results from the proposed experiments will serve as the foundation for comparable experiments in surgically derived breast cancer tissue. These experiments are beyond the time frame of this grant, but we are already stockpiling tissue samples for these future experiments. Our proposed study could provide a new paradigm for hormonal induction of breast cancer via gene amplification, leading to new methods of diagnosis and treatment.

Body: Progress Report (year one)

As described in the DOD funded parent grant, to test our hypothesis we need to map origins of DNA replication in the genome and ask which of these coincide with sites of DNA amplification and with ER binding sites. In the parent grant we proposed to identify potential replication origins by mapping ORC binding sites. However, ORC can also bind to silent origins. We propose to refine this strategy by using newly synthesized (nascent) DNA for direct mapping of all active replication origins in the genome. These data will then be compared to sites of DNA amplification and sites of ER binding to see if a correlation exists.

Task (1) Map replication origins in the human genome. Develop methodology using MCF-7 breast cancer cells to derive a genomic map of replication origins by Helicos sequencing of short nascent strands. To date, replication origins have only been mapped in 1% of the human genome (ENCODE). Our results would be the first replication origin map for the entire human genome.

Subtask (1a) (months 1-6) – preparation of short nascent DNA from MCF-7 cells.

We have **completed** subtask (1a):

The Brodsky lab grew MCF-7 cells to mid-log stage and gave them to the Gerbi lab (postdoc Michael Foulk) for DNA isolation. The nascent strand sequencing work flow is as follows:

Nascent-strand-Seq Work Flow

- Prepare genomic DNA from asynchronous MCF-7 cells with DNAzol (~30% in S-phase)
- Purify Replicative Intermediate (RI) DNA on BND-cellulose (100 ug input)
- Phosphorylate ends with T4-polynucleotide kinase
- Enrich for Nascent Strands by digesting with lambda-exonuclease (λ-exonuclease; Lexo)

- Size select for 1 kb-2 kb nascent strands on low melting point agarose gel to eliminate Okazaki fragments
- Test enrichment of the MYC origin of replication by Real-Time PCR
- Subject nascent strands to fragmentation followed by making them double stranded with random primers and Klenow
- Standard library preparation for Illumina sequencing (200-500bp fragments)
- Sequence library on the Illumina GAIIx platform (pilot experiment) or High-Seq (subsequent experiments) using 42 bp single end reads
- Filter and align reads to the human genome (build hg18) and call peaks (using genomic input DNA as normalization control)

Several labs are now using the methodology we developed for nascent strand sequencing (NS-Seq) whose basis resides in the use of λ-exonuclease to enrich nascent strands coupled with size selection. Our NS-Seq protocol is based on our earlier report (Gerbi and Bielinsky, 1997; Bielinsky and Gerbi, 1998) that nascent DNA is resistant to lambda-exonuclease digestion because of the presence of a 5' RNA primer. This allows the parental DNA to be digested while the nascent DNA is untouched. The nascent strands were size selected on gels for 1-2 kb, which gave greater origin enrichment than a 0.5-1 kb fraction that may have Okazaki fragment contamination (**Figure 1**). Using the c-Myc origin to assess for enrichment, the average of the several preparations used for sequencing had 54-fold enrichment of nascent strands (**Figure 1**). Interestingly, in assessing this enrichment at the c-Myc origin, we discovered that the preferred origin resided in the second exon of the gene while it was previously determined to reside in the promoter of the gene (in HeLa cells: Tao et al., 2000) (**Figure 1**). This observation was confirmed in our NS-Seq data, suggesting plasticity of origin usage at the c-Myc gene in different cell types.

We hope to write an article about the validation of the NS-Seq method and also a Methods article with step by step details of this protocol. The latter will also include a discussion of the computational methods for analysis of the results.

Subtask (1b) (months 7-8) – sequencing and analysis of results to map origins in the genome.
We are close to completion for subtask (1b):
We had proposed that the short nascent DNA preparation would be submitted to the Dana Farber Cancer Institute DNA sequencing facility. They would add poly-A tails to the DNA molecules and carry out Helicos sequencing. They would send us the sequence data which would be analyzed by co-P.I. Ben Raphael to map all active replication origins in the MCF-7 breast cancer genome.

We did one run on the Helicos machine, but there were many errors from the machine and the company has now gone out of business. Therefore, we switched to using the Illumina GAIIX and more recently Illumina Hi-Seq as the platform for sequencing nascent strands. We have obtained the Illumina GAIIX data, have analyzed it, and presented posters on the results at the DOD Era of Hope meeting and the Cold Spring Harbor DNA Replication meetings this summer/earlyfall. Figures from the posters and also the abstracts are attached. The samples have been submitted for Illumina Hi-Seq. The new results will confirm and extend our earlier data from the Illumina GAIIX machine.

Graduate student John Urban identified 53,914 origins in the MCF-7 genome, with a median width of 1.5 kb using the methodology as follows:
We used BEDTools (Quinlan et al., 2010) and features of the genomic analysis of ChIP-Seq data (Euskirchen et al., 2007) for analysis of our data on DNA replication originas in the human genome. 11,805,186 reads of 42 bp were mapped to human genome build hg18 with Bowtie

(Langmead et al, 2009; Langmead 2010). Some of the origins called at loci known to have origin activity are shown via snapshots of the IGV browser. Read coverage and the individual reads were tabulated. IGV tools (Robinson et al, 2011) was used to approximate the fragment coverage (from which single end sequencing reads came) by extending reads to the average fragment length of 350 bp in the direction of the read. Using the Bowtie-mapped reads, MACS (Zhang et al, 2008; Feng et al. 2011) was used to call peaks (called 53,914) by shifting all reads 175 bp in the direction of the read to approximate the center of the average-sized 350 bp fragments, then using the Poisson distribution to call pile-ups enriched over the genomic background coverage with p-value < 0.00001 (--nomodel and --nolambda specified). The breadth of the peak and the summit (the bp/point with highest coverage) were tabulated. The summit is our current approximation for the preferred start site (transition point) of DNA replication for each replication origin in the genome. **Figure 2** presents the results of the statistical analysis of mapping replication origins in the MCF-7 breast cancer genome.

Inter-Origin distances were calculated as those distances between all potential (or most) origins used in a population of MCF-7 cells (**Figure 3**). The maximum distances in our set may be real, but also may be due to weak origins within that space, due to deletions/rearrangements in the cancer genome, or other. The median and mean peak widths both reflect the size of the nascent strands we selected.

The data on the reads for some known origins is presented in **Figure 4.** Many known replication origins were present in our data set including c-Myc (**Figure 4a**), DBF4 (**Figure 4b**), DHFR (**Figure 4c**), β-Globin (**Figure 4d**), RPE (**Figure 4e**), as well as Lamin B2, and Glucose-6-Phosphate Dehydrogenase. There are varying degrees of overlap between our dataset and those of others (Cadoret et al., 2008; Karnani et al., 2010.; Mesner et al. 2011; Martin et al. 2011 – also see Valenzuela et al. 2011), but surprisingly the overlap is less than 40% at best. We did a variety of comparisons between the replication origins reported between each group (**Figure 5**). The Martin and Gerbi origins sets were both pared down to just those origins that lie in the 44 pilot ENCODE regions so that these sets may be directly compared to Cadoret et al. (2008), Karnani et al., and Mesner et al. sets. We took off excess length from each side of the Mesner peaks (assuming origin is centered) to make the relatively large number of overlapped origins (135) more than expectation by random chance. The origin sets of Martin et al (2011), Cadoret et al. and Mesner et al. (2011) are represented in our set by more than random chance, but Karnani's is not. Our comparisons have been summarized as Venn diagram representations of the named set accompanied by a pie chart that represents the named set by breaking it up into degree of overlap of the origins (**Figure 5**) for the current Gerbi ENCODE set (**Figure 5a**), the Martin ENCODE set (**Figure 5b**), the Cadoret set (**Figure 5c**), and the Mesner set (**Figure 5d**). We calculated the statistical significance for the overlap in our data set with the data sets of others for replication origins in the human genome (**Figure 6**); Martin's, Cadoret's, and Mesner's sets are represented in our set by more than random chance, but Karnani's is not.

Also, we constructed a graph representation of pair-wise direct overlaps for the Gerbi set, the Martin set (Aladjem), the Cadoret set (Prioleau), and the Mesner set (Hamlin) where nodes are Sets and weights and directed edges of same color represent how many origins in the set of that color overlap an origin in the set being pointed to (**Figure 7**).

Three of the reports (Cadoret et al., 2008; Karnani et al., 2010.; Mesner et al. 2011) were based on using ENCODE (1% of the human genome) for HeLa cells, so finding only a small amount of overlap could be due to their use of a different cell line than that used by our lab. However, the Martin et al. (2011) report also used MCF-7 cells for NS-Seq of the entire genome and we are puzzled that the agreement was not better between their dataset and ours. They used smaller nascent strands than us and we suspect that may have led to Okazaki fragment contamination in their samples. Indeed, their nascent strand enrichment was less than ours and not even reported in their paper nor were their results validated in their paper. These issues will be dealt with in the discussion of our paper that will be prepared for submission in the coming year. Also, we are collaborating with David Gilbert (University of Florida – Tallahassee) to determine the replication foci higher order structure in the nucleus by chromosome capture methodology and this would also be part of our publication.

To sum up, we are basically on schedule with our experiments, despite the change from the Helicos to the Illumina platform.

Task (2) Comparison of replication origin maps between breast cancer (ER+, ER-) and normal breast cells. These results would indicate if replication origin usage changes between normal and breast cancer cells, and if it varies between ER positive and ER negative breast cancer cells.

We are nearing completion of subtask (2a) to map replication origins in an ER+ breast cancer cell line --- namely MCF-7 (see Task (1). Due to the additional experiments of NS-Seq method validation and collaborative experiments on chromosome capture that were not in the original grant application, the experiments in subtask (2b) to map replication origins in ER- breast cancer cells (e.g., MDAMB231 cells, SKBR3 cells) will be deferred to year two of this grant. Also in year two we will carry out subtask (2c) to map replication origins in normal breast cells (MCF-10A) as in the timeline of the grant application. The similarities and differences in replication origin maps for ER+ and ER- breast cancer cell genomes and comparison to the origin map for the normal breast cell genome will address whether replication origins differ in different cell types, especially comparing breast cancer cells to normal breast cells, and comparing ER+ to ER- breast cancer cells.

Task (3) Correlation of origin map data with sites of (a) DNA amplification and (b) estrogen receptor binding. These data will support or refute the hypothesis that ER may bind next to the replication machinery and induce DNA amplification.

These results that are scheduled for year two will support or refute the hypothesis that ER may bind next to the replication machinery and induce DNA amplification. Co-PI Ben Raphael working together with graduate student John Urban (who has been heavily involved in the computational analysis thus far and presented this work at the Cold Spring Harbor DNA Replication meeting) will correlate origin locations with sites of (a) DNA amplification and (b) estrogen receptor binding as described in **subtask (3a)** of the grant for year two. We will compare the origin map data to data that already exists on sites of DNA amplification (to identify amplification origins) as well as confirm and expand these data using our own data on the number of reads from sequencing bulk genomic DNA from the various cell lines we are using. This information will, in turn, be compared to existing data on sites of ER binding. It may prove necessary to undertake some ChIP (chromatin immunoprecipitation) experiments for validation of ER binding, though not proposed in the original grant application. These data will indicate if a correlation exists between ER binding and origins that re-replicate (amplify), thereby testing our hypothesis.

In the remaining month or two of year two (according to the grant timeline) we hope to engage in pilot runs to map replication origins in surgically derived breast tissue (**subtask (3b)**).We have already begun to stockpile surgically derived breast cancer tissue, provided to us a residual, de-identified tissue from surgeons Theresa Graves and Maureen Chung and pathologist Shamlal Mangray, all from Rhode Island Hospital which is affiliated with the Brown University Medical School. During months 23-24, we will use samples of this tissue to refine the methodology we developed in task (1) for use on surgical specimens. Pilot runs will be initiated in ER + human breast cancer tissue to map replication origins and sites of DNA amplification to compare to matched normal breast tissue from the same patient. These data will be expanded in future studies to reveal if novel origins are used for re-replication and if they correlate with ER binding sites adjacent to them. This information will have clinical importance.

Anticipated papers we hope to publish are:

(1) Methods for Nascent Strand Sequencing (NS-Seq) – molecular biology bench work and computational analysis.
(2) Validation of the Nascent Strand Sequencing method
(3) Identification of replication origins in the MCF-7 human breast cancer genome.
(4) Correlation of replication origins, sites of DNA amplification and estrogen receptor binding

Key Research Accomplishments

- Development of the method of Nascent Strand-Seq (NS-Seq) to map replication origins in the genome - We have developed this method. We plan to also apply it to the yeast genome for validation of the method.

- Application of NS-Seq to map replication origins in the MCF-7 breast cancer genome - We have obtained results of NS-Seq to map replication origins in the MCF-7 genome using the Illumina platform).

- Validation of the NS-Seq results by finding known replication origins in our data set - We have validated NS-Seq on known origins, including Myc, DBF4, DHFR, β-Globin, RPE, as well as Lamin B2, and Glucose-6-Phosphate Dehydrogenase.

- Comparison of our data to the data sets of other labs to map replication origins in the human genome. The data sets from Cadoret et al., 2008; Karnani et al., 2010., and Mesner et al. 2011 were based on using ENCODE (1% of the human genome) for HeLa cells, so finding only a small amount of overlap could be due to their use of a different cell line than that used by our lab. Moreover, even when comparing the results between these three data sets, there was not complete agreement, suggesting lack of saturation of the data. The Martin et al. (2011) data set used MCF-7 cells and was for the full genome, but did not give full overlap with our data. They did not show any data for validation of their results, and we suspect that they had contamination from Okazaki fragments as they selected small nascent strand DNA.

Reportable Outcomes

Our results have been presented at the DOD Era of Hope meeting and the Cold Spring Harbor DNA Replication meeting (see attached abstracts) and will be written up for publication soon.

(1) Gerbi SA, Foulk, M, Brodsky A and Raphael B (2011). Origins of DNA Replication and Amplification in the Breast Cancer Genome. Department of Defense Breast Cancer Research Program Era of Hope meeting (August 2-5, 2011; Orlando, Florida) p. 69 (poster abstract 48-8).

(2) Urban J, Foulk M, Casella C and Gerbi SA (2011). Mapping DNA replication origins to the human genome. Poster presentation at the Cold Spring Harbor Laboratory meeting on Eukaryotic DNA Replication and Genome Maintenance (September 6-10, 2011; Cold Spring Harbor, NY).

Conclusion

We are on schedule according to the timetable in our grant application, and are excited by our results mapping all replication origins in the human genome of MCF-7 breast cancer cells.

Personnel Paid From This Grant

<u>PI and co-PIs:</u>

Susan Gerbi	Professor of Biology (PI)
Alexander Brodsky	Assistant Professor of Medical Science (co-PI)
Benjamin Raphael	Associate Professor of Computer Science (co-PI)

<u>Lab Personnel:</u>

Jacob Bliss	Research Assistant
Stephen Doris	Postdoctoral Research Associate
Elijah Douglass	Research Assistant
Michael Foulk	Research Associate
Yutaka Yamamoto	Research Associate

References Cited

AK Bielinsky AK and Gerbi SA (1998). Discrete start sites for DNA synthesis in the yeast ARS1 origin. Science 279:95-98.

JC Cadoret JC, Meisch F, Hassan-Zadeh V, Luyten I, Guillet C, Duret L, Quesneville H and Prioleau MN (2008). Genome-wide studies highlight indirect links between human replication origins and gene regulation. Proc. Nat. Acad. Sci. 105: 15837-15842.

Euskirchen GM, Rozowsky JS, Wei CL, Lee WH, Zhang ZD, Hartman S, Emanuelsson O, Stolc V, Weissman S, Gerstein MB, Ruan Y, Snyder M. (2007) Mapping of transcription factor binding regions in mammalian cells by ChIP: comparison of array- and sequencing-based technologies. Genome Res. 17: 898-909.

Feng J, Liu T and Zhang Y. (2011). Using MACS to identify peaks from ChIP-Seq data. Curr Protoc Bioinformatics, Chapter 2:Unit 2.14.

Gerbi SA and Bielinsky AK (1997). Replication initiation point mapping. Methods 13 (3): 271-280.

Karnani N, Taylor CM, Malhotra A, Dutta A. (2010) Genomic study of replication initiation in human chromosomes reveals the influence of transcription regulation and chromatin structure on origin selection. Mol Biol Cell. 21: 393-404.

Langmead B, Trapnell C, Pop M and Salzberg SL. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol.10(3):R25.

Langmead B. (2010). Aligning short sequencing reads with Bowtie. Curr Protoc Bioinformatics. Chapter 11:Unit 11.7

Martin MM, Ryan M, Kim R, Zakas AL, Fu H, Lin CM, Reinhold WC, Davis SR, Bilke S, Liu H, Doroshow JH, Reimers MA, Valenzuela MS, Pommier Y, Meltzer PS and Aladjem MI (2011).

Genome-wide depletion of replication initiation events in highly transcribed regions. Genome Res. (Sept. 22, 2011 Epub).

Mesner LD, Valsakumar V, Karnani N, Dutta A, Hamlin JL and Bekiranov S. (2011). Bubble-chip analysis of human origin distributions demonstrates on a genomic scale significant clustering into zones and significant association with transcription. Genome Res 21:377-389.

Neve RM, Chin K, Fridlyand J, Yeh J, Baehner FL, Fevr T, Clark L, Bayani N, Coppe JP, Tong F, Speed T, Spellman PT, DeVries S, Lapuk A, Wang NJ, Kuo WL, Stilwell JL, Pinkel D, Albertson DG, Waldman FM, McCormick F, Dickson RB, Johnson MD, Lippman M, Ethier S, Gazdar A and Gray JW (2006). A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell* 10: 515-527.

Quinlan AR, Hall IM. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26: 841-2.

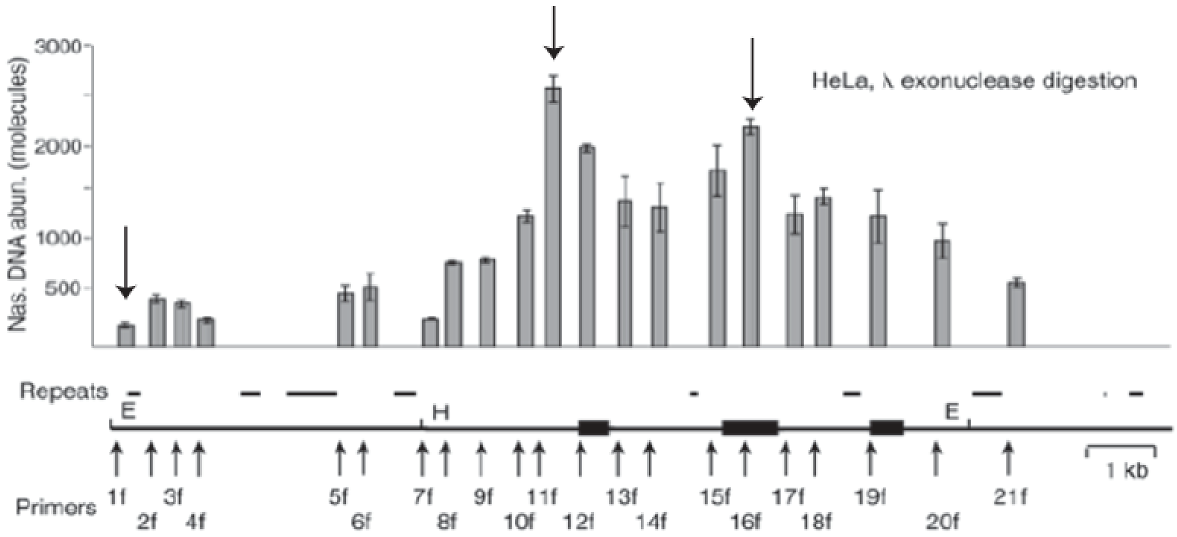
Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz and Mesirov JP (2011).Integrative Genomics Viewer. Nature Biotech 29: 24–26.

Tao L, Dong Z, Leffak M, Zannis-Hadjopoulos M and Price G (2000). Major DNA replication initiation sites in the c-myc locus in human cells. J. Cell Biochem 78:442-457

Valenzuela MS, Chen Y, Davis S, Yang F, Walker RL, Bilke S, Lueders J, Martin MM, Aladjem MI, Massion PP and Meltzer PS. (2011). Preferential localization of human origins of DNA replication at the 5'-ends of expressed genes and at evolutionarily conserved DNA sequences. PLoS One. 2011;6(5):e17308.

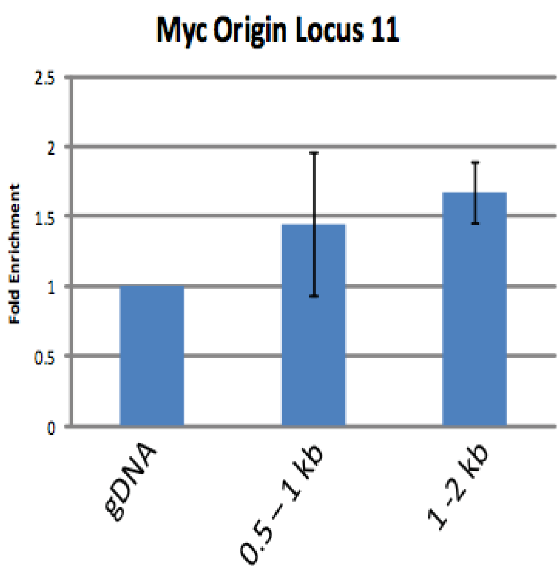
Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W and Liu XS. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol.9(9):R137.

Figure1a: Map of the Myc Locus, showing locations of primer pairs. Primer pair 11 is in front of the first exon; primer pair 16 is within the second exon, and both show strong origin activity.



From Tao et al., JCB 78:442-57 (2000)

Figure 1b: Origin enrichment in short nascent DNA strand preparations determined by real time PCR. We found that locus 16 (in the second exon) had more origin activity than locus 11 (before exon 1). Moreover, the 1-2 kb size fraction showed a greater enrichment for origin activity than the 0.5=1 kb size fraction, perhaps due to some contamination by Okazaki fragments in the latter.



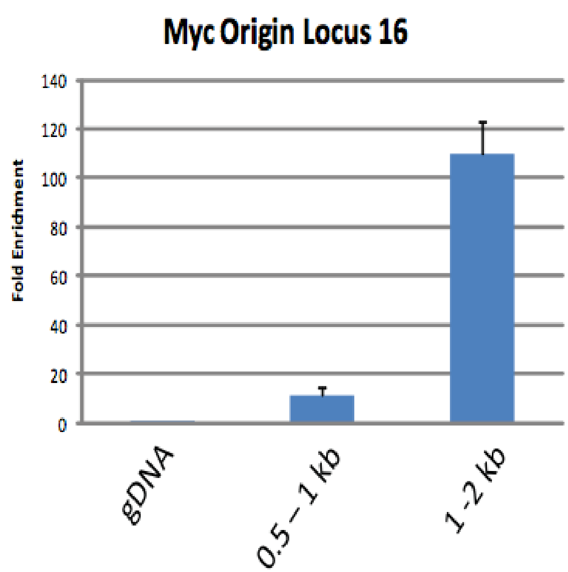


Figure 2: Statistics on MCF-7 Breast Cancer Cell Replication Origin Mapping.

Trial3_Peak_statistics.txt

Last Saved: 8/12/11 9:26:10 AM
File Path ▾ : ~/Gerbi_Lab/Trial3_Peak_statistics.txt

```
Without Y chromosome, there are 53,914 peaks

mean(peak length) = 1,915.6 = ~1,916
median(peak length) = 1,514
max(peak length) = 137,903
min(peak length) = 381
stdev(peak length) = 2,158.9 = ~ 2,159

numberpeaks>mean = 14,559
numberpeaks<mean = 39,355
numberpeaks=mean = 0
Biggest 10 sizes:
63524, 66920,68762,71122,76923,84172,88529,92529,127624,137903

10 smallest sizes:
933, 1236,1336,1396,1462, 1595, 1941, 2407, 2658, 8625
```

chrs_q_arms_statistics.xlsx

chr#	~size_armQ	~mappable_si	#origins_arm(	mean_midpoir	median_midpt	max_midpt_di	min_midpt_di	stdev_midpt_(	#interdist>me	#interdist<me	total # distanc
chr1			2,173	48,623	11,063	798,123	1,075	88,666	558	1,614	2,172
chr2			2,108	70,168	19,144	1,339,600	992	121,030	588	1,519	2,107
chr3			1,687	61,937	24,332	663,910	1,045	89,928	510	1,176	1,686
chr4			972	143,050	90,796	1,273,300	1,478	160,940	336	635	971
chr5			1,926	68,234	22,680	924,780	1,142	107,300	551	1,374	1,925
chr6			851	128,140	71,788	1,725,400	916	159,840	284	566	850
chr7			1,840	53,140	16,099	863,160	1,048	91,787	497	1,342	1,839
chr8			2,266	43,843	12,384	757,960	1,179	73,258	609	1,656	2,265
chr9			2,524	29,700	7,156	967,336	888	66,950	512	2,011	2,523
ch10			1,390	67,417	19,224	1,394,546	1,057	122,490	373	1,013	1,389
chr11			1,922	41,616	7,500	1,165,825	1,092	100,270	373	1,548	1,922
chr12			1,813	52,878	14,919	890,070	1,064	94,234	473	1,339	1,812
chr13			681	141,190	58,340	1,525,121	1,055	195,940	210	470	680
chr14			2,096	41,944	10,733	675,242	905	73,331	544	1,551	2,095
chr15			1,952	42,063	11,728	1,209,300	904	84,543	482	1,469	1,951
chr16			1,705	25,698	6,254	939,200	1,159	70,591	289	1,415	1,704
chr17			3,081	18,297	6,107	821,815	927	45,865	619	2,461	3,080
chr18			312	190,700	96,245	1,940,854	1,874	246,740	103	208	311
chr19			1,380	22,742	8,931	636,736	1,035	44,936	314	1,065	1,379
chr20			2,327	14,784	5,699	1,002,615	1,154	40,960	398	1,928	2,326
chr21			1,021	36,443	6,050	3,050,856	1,350	150,450	126	894	1,020
chr22			762	46,135	16,300	909,230	1,049	85,206	200	561	761
chrX			916	101,950	47,221	1,276,059	1,113	139,130	298	617	915
chrY?											

chr#	~size_armP	~mappable_si	#origins_armI	mean_midpoir	median_midpt	max_midpt_di	min_midpt_di	stdev_midpt_(	#interdist>me	#interdist<me	total # distanc
chr1	-	-	2019	60,051	18,285	1,253,279	1138	114,380	526	1,492	2018
chr2			1635	56,024	16,309	1,012,652	1,141	97,374	431	1,203	1634
chr3			2041	44,298	8,743	1,004,000	1,018	88,381	484	1,556	2040
chr4			623	79,325	18,231	1,176,197	1199	154,080	155	467	622
chr5			810	57,310	15,190	1,294,000	1,043	112,400	215	594	809
chr6			977	60,298	21,596	1,187,200	1,022	98,550	270	706	976
chr7			1396	41,549	8,442	919,310	1,063	89,334	318	1,077	1395
chr8			361	122,020	46,769	1,376,900	1,225	187,720	106	254	360
chr9			382	123,510	62,304	1,713,400	1,218	185,640	121	260	381
ch10			603	64,992	16,838	896,550	1,198	110,900	170	430	602
chr11			620	83,003	28,254	1,025,518	1,393	131,870	165	454	619
chr12			383	90,894	39,012	1,166,011	1,555	145,500	110	272	382
chr13	13p not mappable (NNNN)		NA	NA	NA	NA	NA	NA	NA	NA	NA
chr14	14p not mappable (NNNN)		NA	NA	NA	NA	NA	NA	NA	NA	NA
chr15	15p not mappable (NNNN)		NA	NA	NA	NA	NA	NA	NA	NA	NA
chr16			1,648	21,335	6,625	819,340	1,065	55,546	310	1,337	1,647
chr17			698	31,968	10,406	625,830	1,077	62,246	166	531	697
chr18			152	101,460	53,384	680,340	1,414	121,200	48	103	151
chr19			1,101	22,188	8,732	458,350	1,368	39,604	272	828	1100
chr20			394	66,678	20,115	1,011,693	1,031	114,690	100	293	393
chr21	21p not		NA	NA	NA	NA	NA		NA	NA	NA
chr22	22p not		NA	NA	NA	NA	NA		NA	NA	NA
chrX			366	210,170	86,002	1,610,900	1,416		112	253	365
chrY?											

Figure 3: Inter-Origin distances here are those distances between all potential (or most) origins used in a population of MCF-7 cells. The max distances in our set may be real, but also may be due to weak origins within that space, due to deletions/rearrangements in the cancer genome, or other. The median and mean peak widths both reflect the size of the nascent strands we selected.

	Inter-Origin Distance Statistics				Nascent-Strand Peak Sizes	
# putative origins	Median	Mean	Min	Max	Median	Mean
53,914	12,116 bp	52,878 bp	887 bp	3,050,886 bp	1,514 bp	1,916 bp

Figure 4: 11,805,186 reads of 42-bp were mapped to human genome build hg18 with Bowtie (Langmead et al, 2009). Some of the origins called at loci known to have origin activity are shown via snapshots of the IGV browser. In each, the 2nd/3rd rows show read coverage and the individual reads respectively. Note that in some of the read pile-ups, many reads are not seen in this freeze-frame (scrolling in the IGV browser is necessary). IGVtools (Robinson et al, 2011) was used to approximate the fragment coverage (from which single end sequencing reads came) by extending reads to the average fragment length of 350bp in the direction of the read (top row). Using the Bowtie-mapped reads, MACS (Zhang et al, 2008) was used to call peaks (called 53,914) by shifting all reads 175bp in the direction of the read to approximate the center of the average-sized 350bp fragments, then using the poisson distribution to call pile-ups enriched over the genomic background coverage with p-value < 0.00001 (--nomodel and --nolambda specified). The 4th row shows the breadth of the peak while the 5th (bottom-most) row shows the summit (the bp/point with highest coverage). The summit is our current approximation for the preferred start site (transition point).

Figure 4a - Myc Locus:

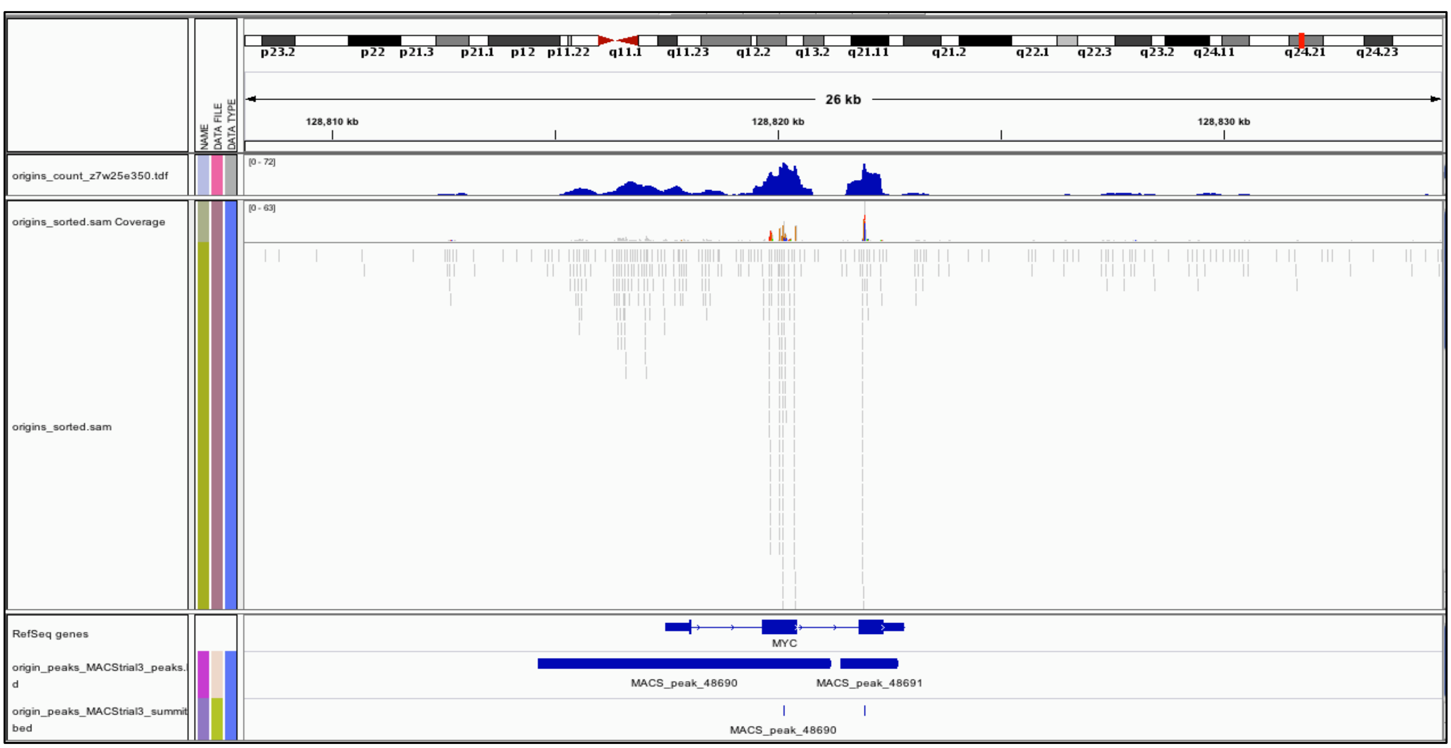


Figure 4b - DBF4 Locus:

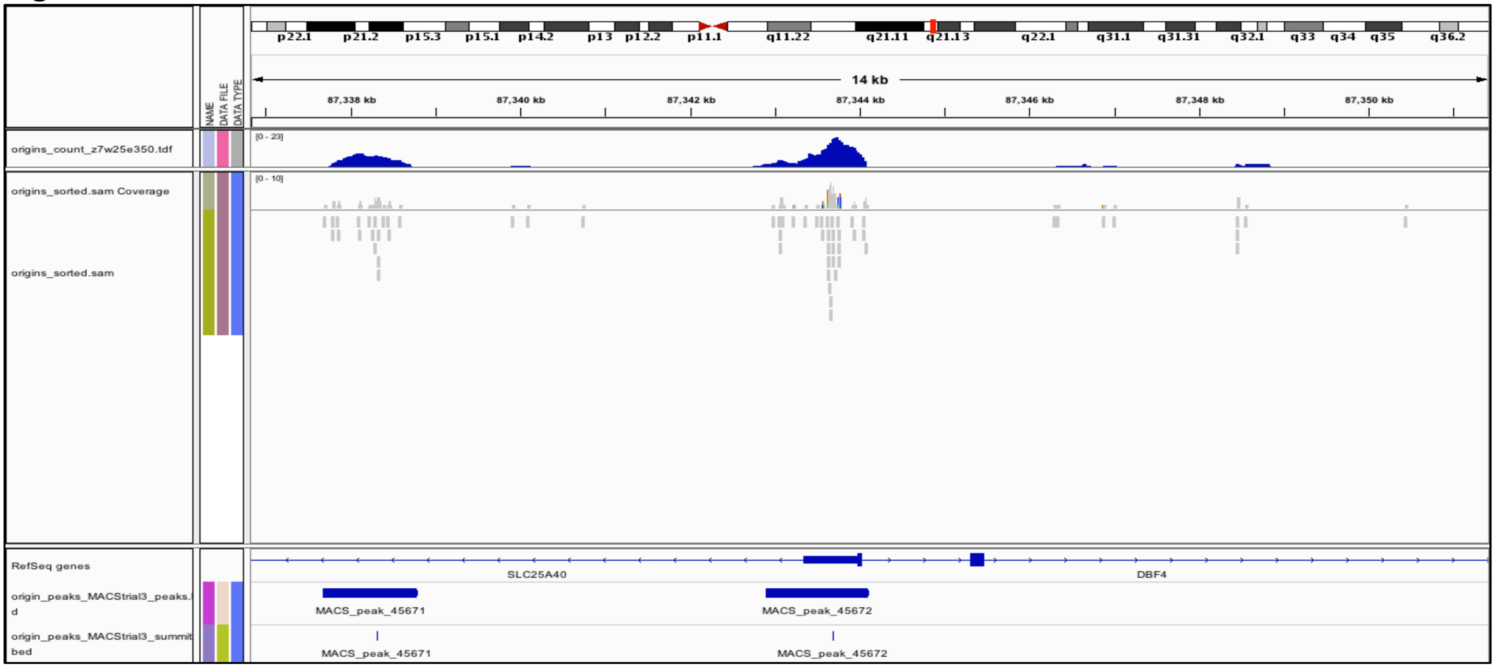


Figure 4c - DHFR Locus:

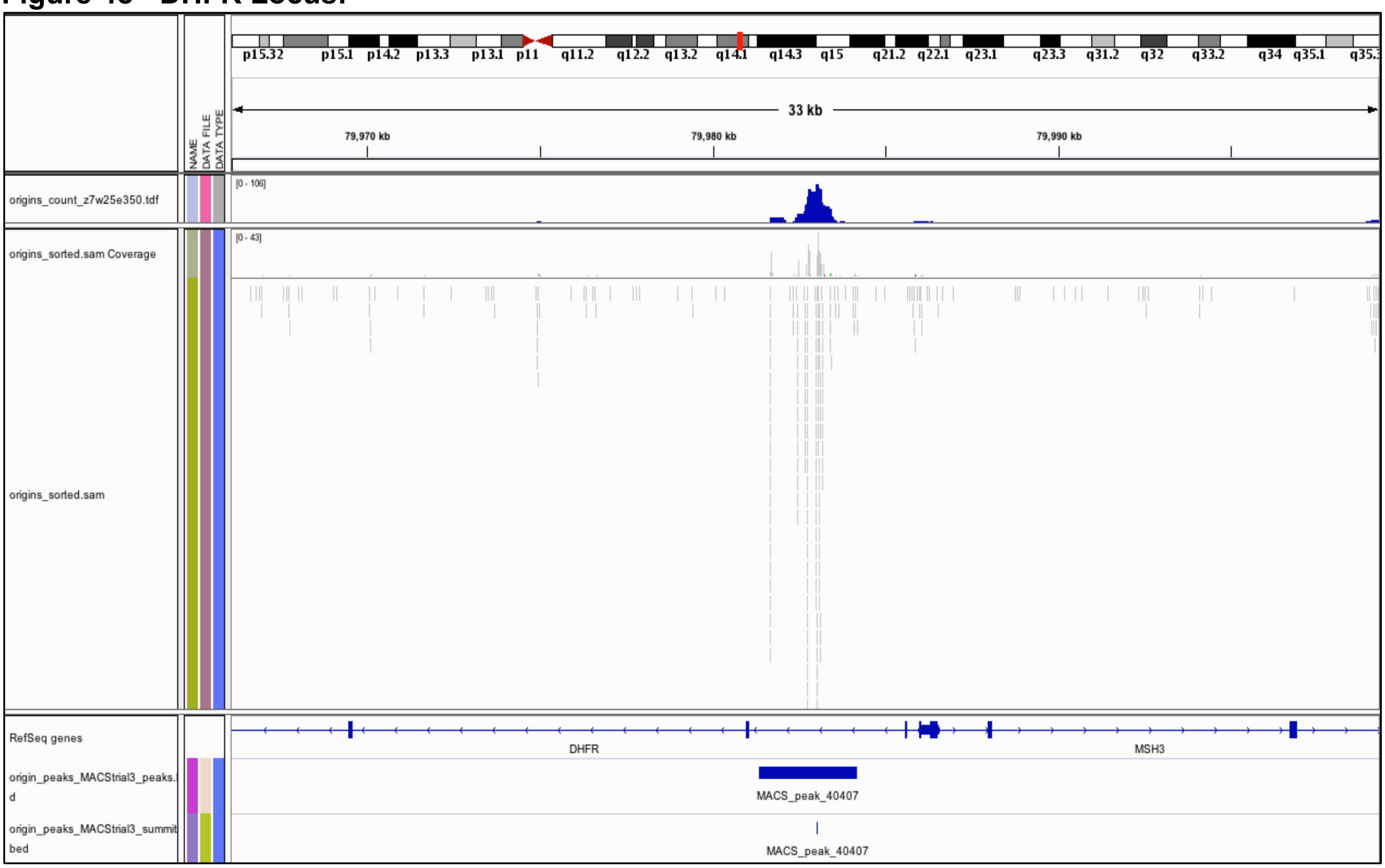


Figure 4d - β-globin Locus:

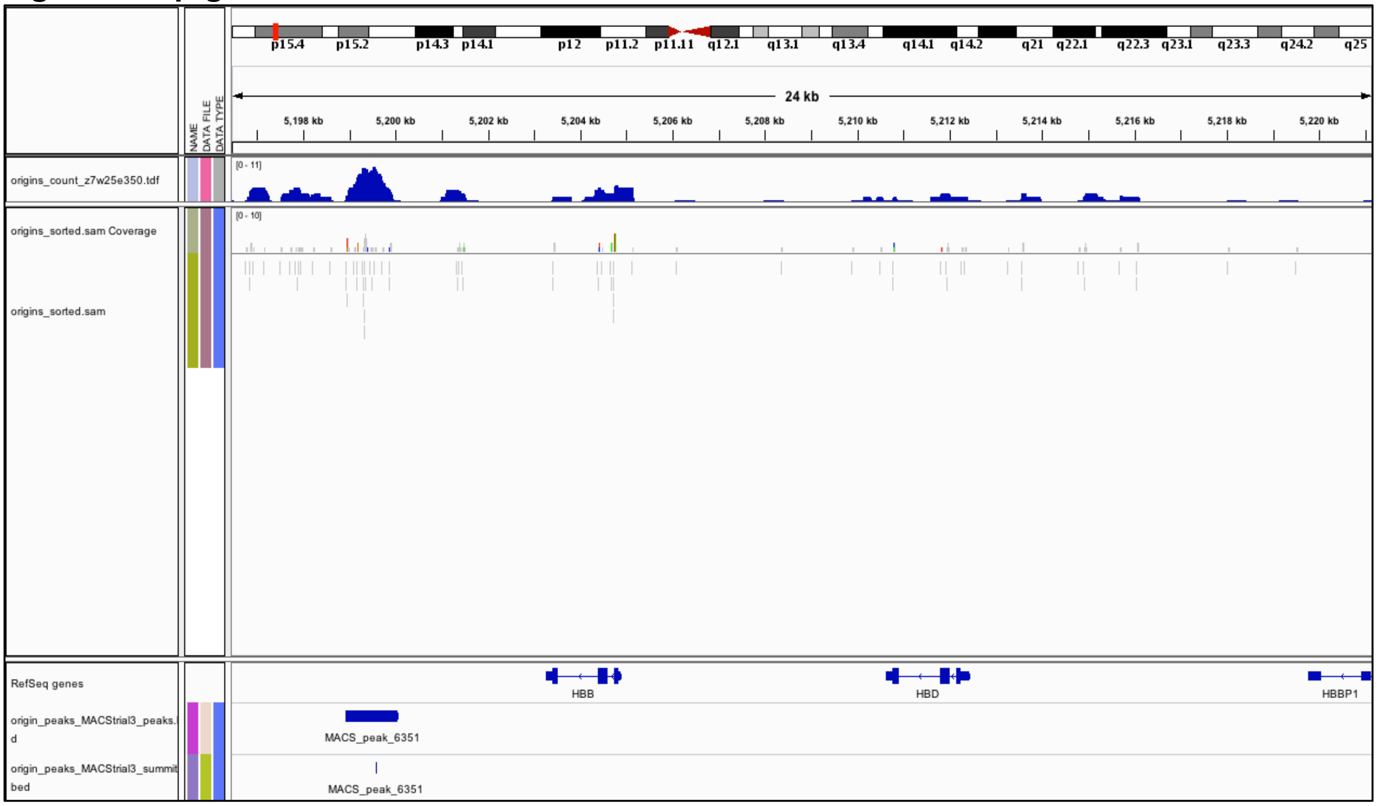


Figure 4e - RPE Locus:

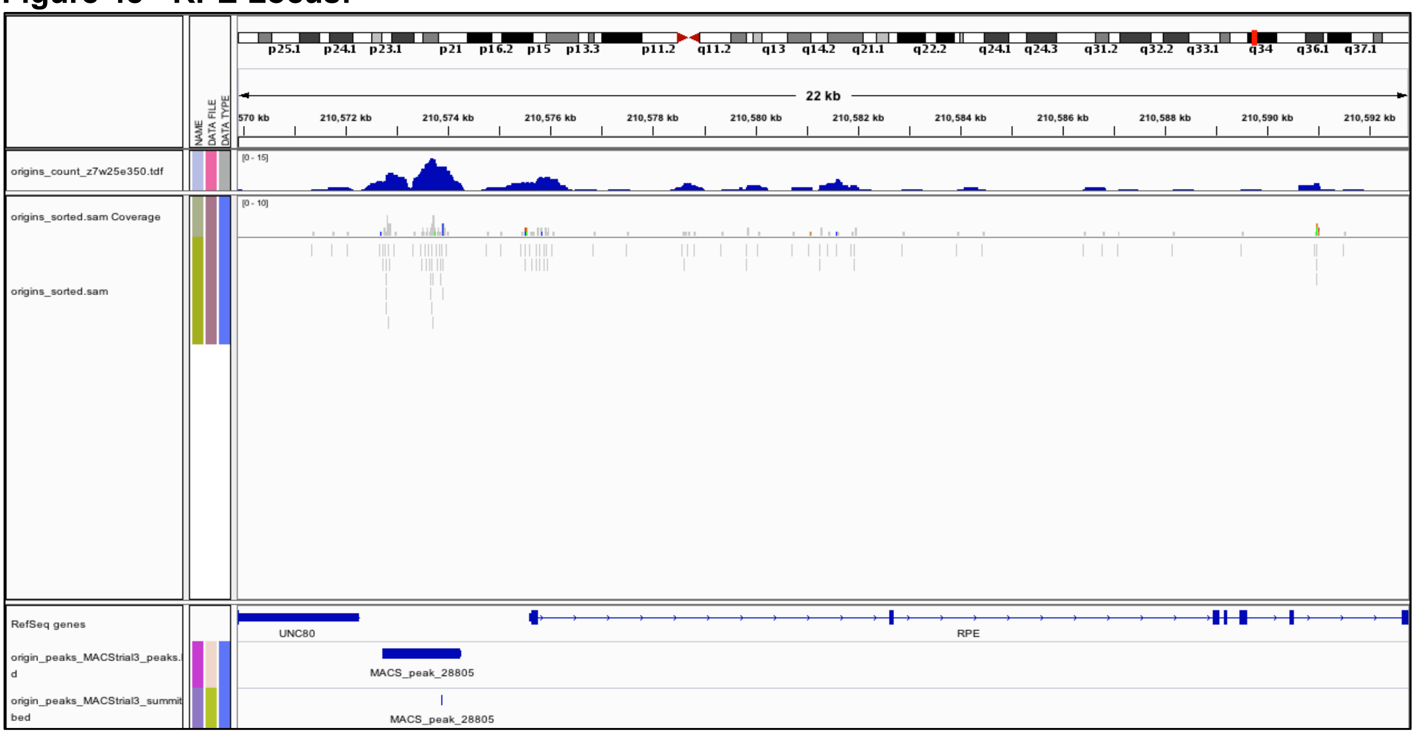
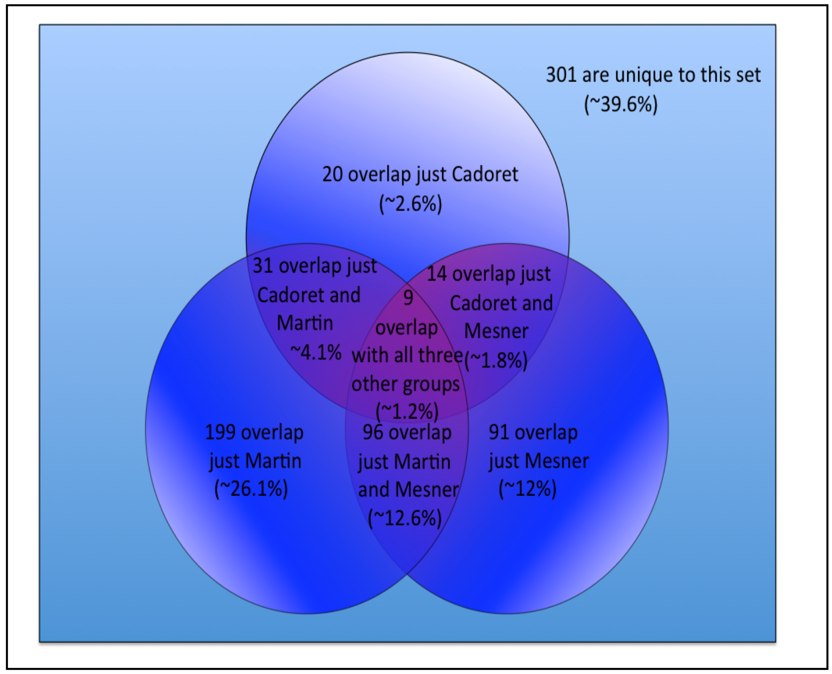


Figure 5: Pictorial representations of DNA replication origin sets

Figure 5a - Gerbi
The current Gerbi Origin Set contains 761 in 44 pilot ENCODE regions



- Unique (found in this set only)
- Origins in this set with overlap in one other set
- Origins in this set with overlap in 2 other sets
- Origins in this set with overlap in 3 other sets

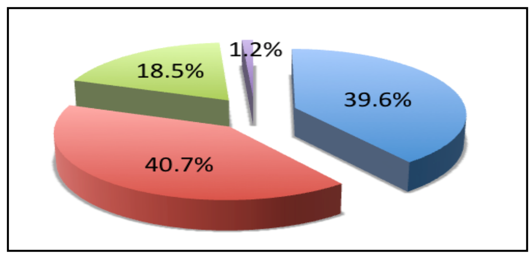
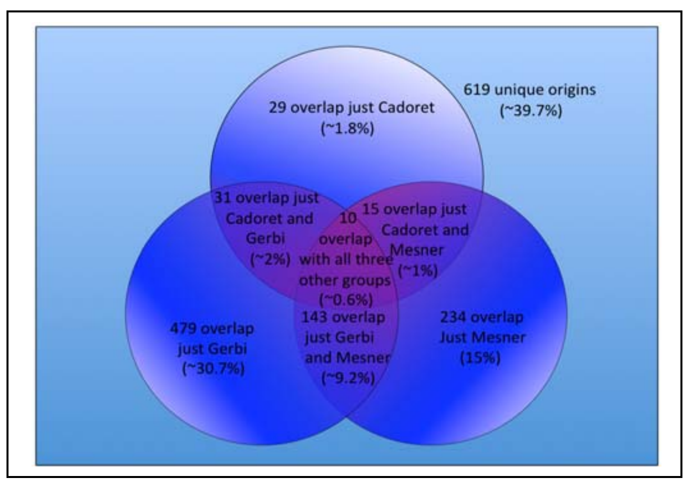


Figure 5b - Martin
The Martin Origin Set contains 1,560 in 44 pilot ENCODE regions



- Unique (found in this set only)
- Origins in this set with overlap in one other set
- Origins in this set with overlap in 2 other sets
- Origins in this set with overlap in 3 other sets

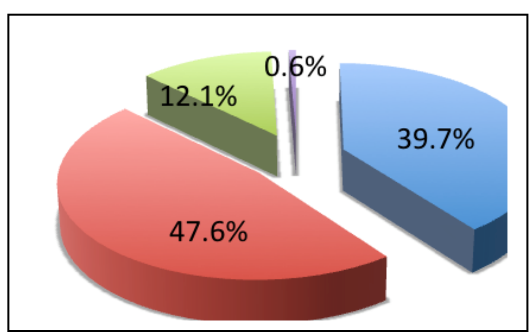
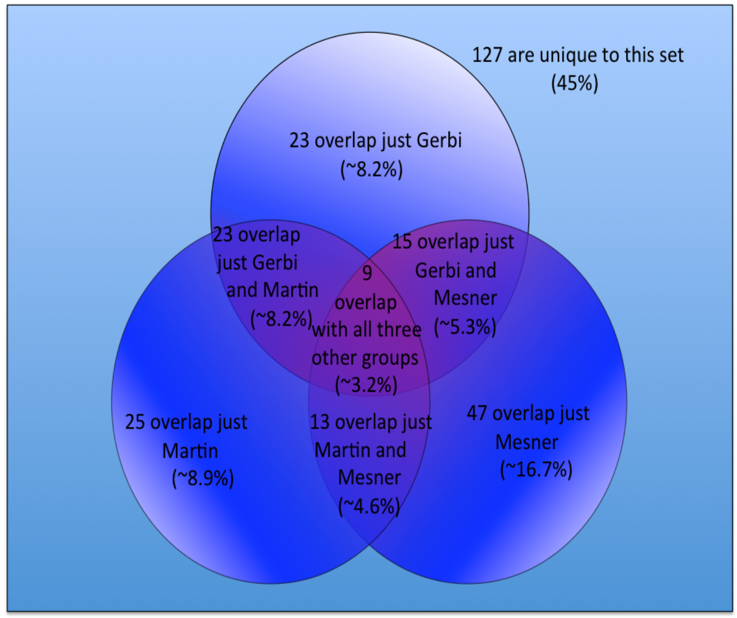


Figure 5c - Cadoret
The Cadoret Origin Set Contains 282 in 44 pilot ENCODE regions after lift-Over to hg18 (283 in hg17)

- Unique (found in this set only)
- Origins in this set with overlap in one other set
- Origins in this set with overlap in 2 other sets
- Origins in this set with overlap in 3 other sets



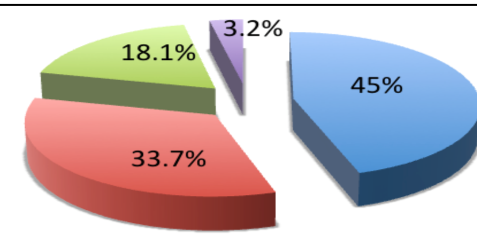
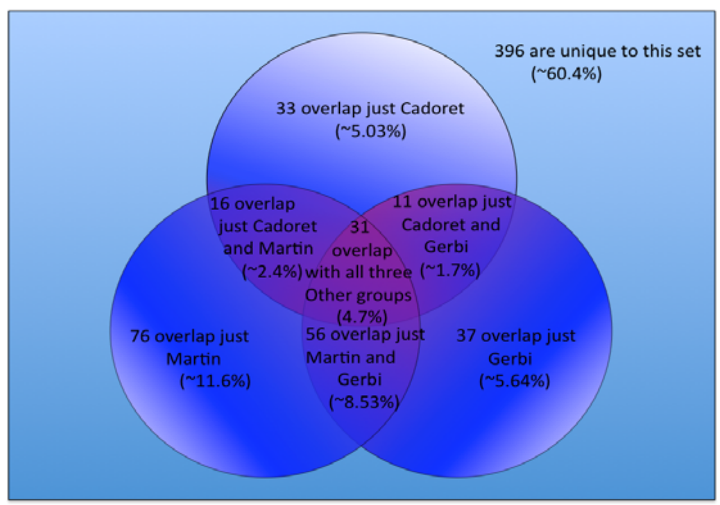


Figure 5d - Mesner
The Mesner Origin Set Contains 656 in 44 pilot ENCODE regions after lift-Over to hg18 (283 in hg17)

- Unique (found in this set only)
- Origins in this set with overlap in one other set
- Origins in this set with overlap in 2 other sets
- Origins in this set with overlap in 3 other sets



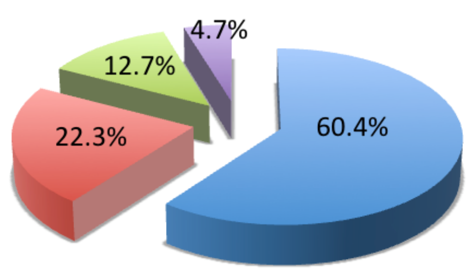
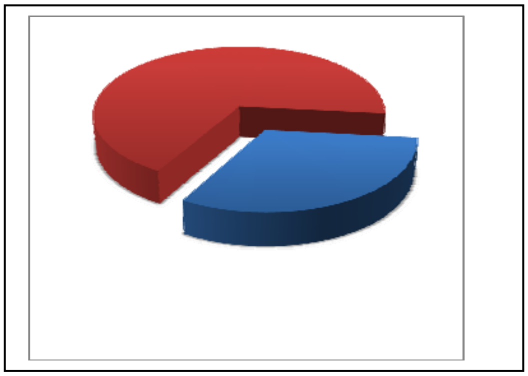


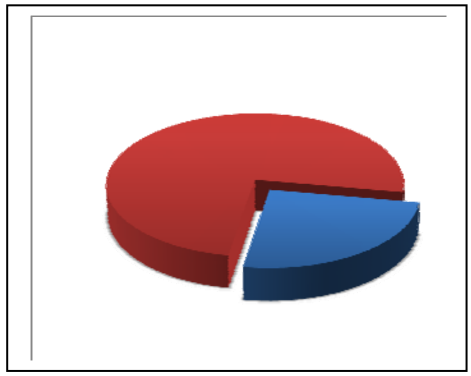
Figure 6: Statistics of replication origin overlap between the Gerbi data set and other data sets

Martin



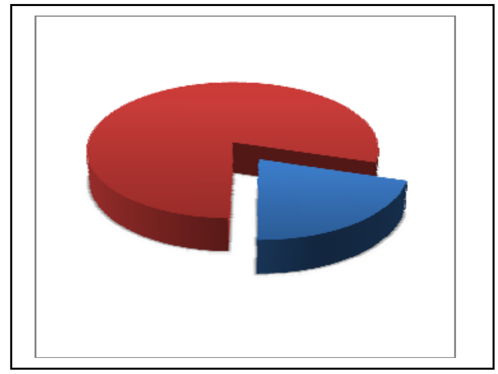
Expect:~5972/93905
= ~6.36%
Observe: 29624/93905
= ~31.55%
p-value: <<<10^-20

Cadoret



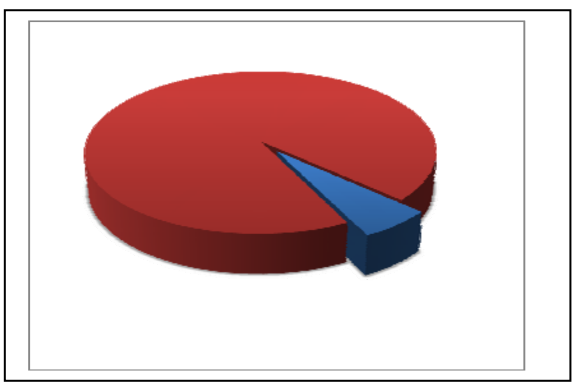
Expect: ~26/282
= ~9.2%
Observe: 70/282
= ~31.55%
p-value: 3.77x10^-15

Mesner (bubble)



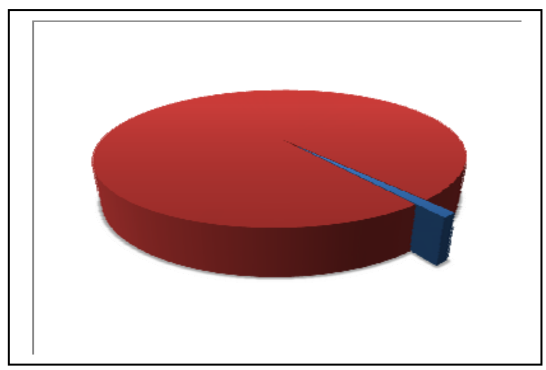
Expect: ~145.6/656
= ~22.2%
Observe: 135/656
= 20.6%
p-value: ~0.83**

Karnani (Lexo)



e Expect: ~28.5/320
= ~8.9%
Observe: 19/320
= ~5.9%
p-value: ~0.966

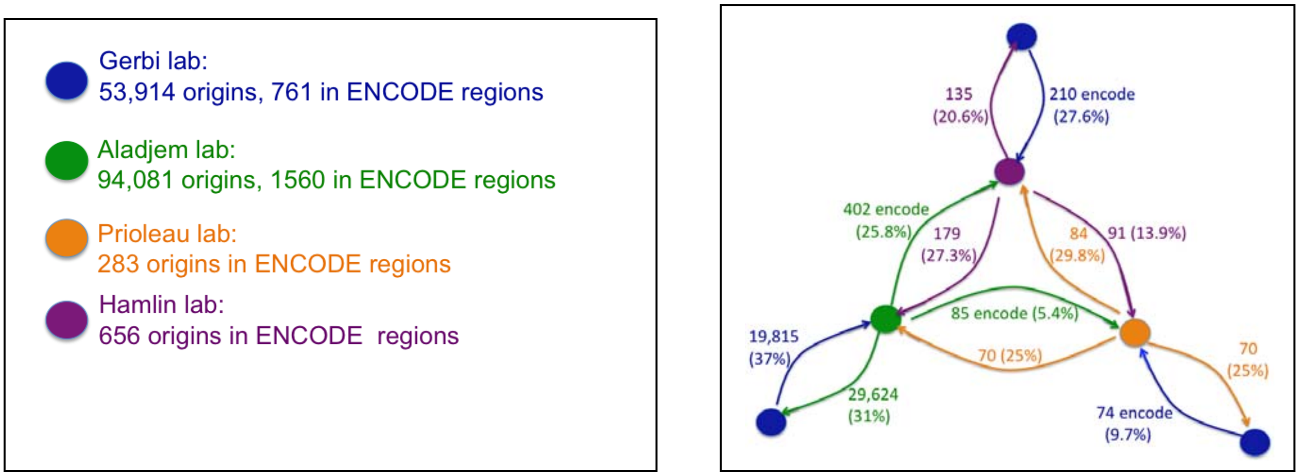
Karnani (BrdU-IP)



Expect: ~ 70.8/761
= ~9.3%
Observe: 9/761
= ~1.18%
p-value: ~1

The Mesner data is not more than random chance probably due to the large size of peaks (~3.5-4.5x bigger than most other sets). Taking off excess length from each side of the Mesner peaks (assuming origin is centered) may make the relatively large number of overlapped origins (135) more than expectation by random chance. Martin's, Cadoret's, and Mesner's sets are represented in our set by more than random chance, but Karnani's is not.

Figure 7: A graph representation of pair-wise direct overlaps for the Gerbi set, the Martin set (Aladjem), the Cadoret set (Prioleau), and the Mesner set (Hamlin). Nodes are Sets. Weights and directed edges of same color represent how many origins in the set of that color overlap an origin in the set being pointed to.



Abstract for the DOD *Era of Hope* meeting, August 2-5, 2011 (Orlando, FL), # p48-8

"Origins of DNA Replication and Amplification in the Breast Cancer Genome"

Susan A. Gerbi (Presenter), Michael Foulk, Alexander Brodsky and Benjamin Raphael.
Brown University, Providence, RI 02912

The fidelity of DNA replication is of paramount importance for normal function of a cell. Disregulation of replication can lead to DNA amplification that is a hallmark of cancer. When oncogenes are amplified, they promote growth of the cancerous cell. Hence, it is important to understand the mechanism underlying DNA amplification. We suggest that re-firing of an origin of replication may be an initiating event in DNA amplification. Our previous research on developmentally regulated DNA amplification in a model system of the fly *Sciara* demonstrated that a steroid hormone triggers re-firing of a DNA replication origin, resulting in DNA amplification. The steroid hormone estrogen has been implicated in breast cancer progression. Can our previous results in the fly serve as a paradigm --- can estrogen induce DNA amplification in breast cancer? We want to learn whether binding sites for the estrogen receptor are located adjacent to origins of DNA amplification in the genome of MCF-7 breast cancer cells. Sites of DNA amplification and sites of binding of the estrogen receptor have already been identified in the MCF-7 genome. To map the origins of DNA amplification requires that we map all replication origins in the MCF-7 breast cancer genome. In order to identify origins of replication, we have made preparations of short nascent strands that will be sequenced using next generation sequencing technology. In brief, nascent DNA is resistant to lambda-exonuclease digestion because of the presence of a 5' RNA primer, allowing the parental DNA to be digested while the nascent DNA is untouched. In a preliminary experiment we were able to enrich (up to 19-fold) for nascent strands from asynchronously growing MCF-7 cells that were subsequently sequenced by Illumina. Our data overlapped with 78 of the 283 replication origins identified in HeLa cells by Cadoret et al. (2008) in the ENCODE region of the human genome. Encouraged by these results, we optimized the protocol to further enrich for nascent DNA, adding precautions to stabilize the RNA primer on the nascent DNA. Using the c-Myc origin to assess for enrichment, we have produced several preparations with substantial enrichment (up to 100-fold). We have sent these nascent strands for sequencing using Helicos single molecule sequencing and are currently analyzing the data. We also intend to use Illumina to sequence nascent strands in the near future and compare the results between the two platforms. Mapping all the replication origins in the MCF-7 genome will allow us to identify which origins occur at regions of DNA amplification and whether they reside in close proximity to estrogen receptor binding sites.

(Supported by DOD CDMRP log # BC097936)

Abstract for the Cold Spring Harbor Laboratory meeting on *Eukaryotic DNA Replication and Genome Maintenance*, September 6-10, 2011, Cold Spring Harbor, NY., poster 223.

MAPPING DNA REPLICATION ORIGINS TO THE HUMAN GENOME

John Urban*, Michael Foulk*, Cinzia Casella and Susan A. Gerbi. Brown University BioMed Division, Providence, RI 02912 USA (* co-first authors)

We have mapped replication origins in the human genome using next generation sequencing technology. Asynchronous MCF-7 human breast cancer cells in log phase were used for preparations of short nascent strands for sequencing on the Illumina platform. Our nascent strand-seq ("NS-Seq") protocol is based on our earlier report (AK Bielinsky & SA Gerbi. 1998. Science 279:95-8) that nascent DNA is resistant to lambda-exonuclease digestion because of the presence of a 5' RNA primer. This allows the parental DNA to be digested while the nascent DNA is untouched. The nascent strands were size selected on gels for 1-2 kb, which gave greater origin enrichment than a 0.5-1 kb fraction that may have Okazaki fragment contamination. Using the c-Myc origin to assess for enrichment, the average of the several preparations used for sequencing had 54-fold enrichment of nascent strands. Interestingly, in assessing this enrichment at the c-Myc origin, we discovered that the preferred origin resided in the second exon of the gene while it was previously determined to reside in the promoter of the gene (in HeLa cells: L Tao et al. 2000. J. Cell Biochem 78:442-57). This observation was confirmed in our NS-Seq data, suggesting plasticity of origin usage at the c-Myc gene in different cell types. We identified 53,914 origins in the MCF-7 genome, with a median width of 1.5 kb. Many known replication origins were present in our data set including c-Myc, DHFR, Dbf4, Lamin B2, beta-Globin and Glucose-6-Phosphate Dehydrogenase. There are varying degrees of overlap between our dataset and those of others (JC Cadoret. 2008. PNAS 105:15837-42; N Karnani et al. 2010. Mol Biol Cell 21:393-404; Mesner et al. 2011. Genome Res 21:377-89; MM Martin et al. 2011. Genome Res) as will be discussed.
(Supported by DOD CDMRP log # BC097936)